



TITLE:

Policy Improvement in Markov Decision Processes and Markov Potential Theory (数理計画と決定過程論)

AUTHOR(S):

安田, 正実

CITATION:

安田, 正実. Policy Improvement in Markov Decision Processes and Markov Potential Theory (数理計画と決定過程論). 数理解析研究所講究録 1977, 299: 93-113

ISSUE DATE:

1977-06

URL:

<http://hdl.handle.net/2433/106269>

RIGHT:

Policy improvement in Markov Decision Processes and Markov potential theory

千葉大 教養部 安田正実

1. Introduction

Markov Decision Process (MDP)

と Markov potential theory との関係は、直接的なもの、すなわち MDP における potential theory と間接的なもの、potential theory の結果による MDP のそのの別証明とに分類される。後者に属するものとして Shaufele [11] の論文があるが、前者の方が我々計画法を研究するものにとって興味深い。これについては、まだ理論体系が形成されていない。しかし、渡辺若 [14] は LP の dual problem と関連して、Howard [7] の iteration ^{における} 単調性の意味を potential の用語で説明したし、N. Furukawa [6] や H. Aso and M. Kimura [1] は potential kernel の性質を用い、policy improvement を証明した。また A. Hordijk [8] も MDP をその定式化から potential の議論として構成している。これらはその理論形成を意図しているように思われる。

しかし、多くの場合、解析的なとり扱いが容易なことから transient chain に対応する case に制限されている。そこで、この報告では、MDP の policy improvement に役立つよう、一般の chain に対して新しい potential を定義する。そしてこの potential を用いて MDP の policy improvement の解釈とその証明を与えるのが目的である。

まず 解釈は「MDP の reward の増加は improvement における increment を charge とする potential と regular function の和で表わされる」という形である。Transient chain のとき、potential は従来のそれと一致し、regular function は 0 であって 渡辺浩 [14] の与えたものに帰着される。また 各種の improvement の証明も denumerable state space であっても可能で、その方法の本質はすべてに通じて同一であることがわかる。ここで考える MDP の policy improvement は次の場合；

- (1) discounted case (2) average case
- (3) nearly optimal case
- (4) sensitive discounted case

このうち (1), (2) は MDP の代表的なものであって数多くの文献がある。とくに (1) について Howard [7], Blackwell [2], [3], (2) について Howard [7], Derman

[4], [5] を挙げておく。Blackwell [2] は (1) と同時に (3) もその存在性を議論している。その後これを糸口として発展した (4) は Miller and Veinott [10], Veinott [12]; [13] に依るものである。

2. Potential theory S を finite または

denumerable な自然数の部分集合, P を $S \times S$ 上の Markov 行列, P^* を P のベキ $\{P^n\}$ の cesaro 和

$$P^* = \lim_{k \rightarrow \infty} \frac{1}{k} \{I + P + \dots + P^{k-1}\} \quad (2.1)$$

I は単位行列で \lim は point wise の収束。この P^* の存在は知られていて 次の性質をもつ。

$$0 \leq P^* = P^*P = P^*P = P^*P^*, \quad P^*\underline{1} \leq \underline{1} \quad (2.2)$$

ここで $\underline{1}$ は恒等的に 1 とする関数またはベクトルとする。

\leq は成分毎で, §3 にて用いる " \leq " の定義は各行ベクトル^{の差}について最初に 0 でない要素があれば, それが正のときと定める。関数 f が P について regular とは

$Pf = f$ のとき。Markov chain についての定義などは Kemeny, Snell and Knapp [9] を参照されたい。

定義 1 S 上の関数 f が Markov 行列 P に関する charge であるとは

$$H_n = \sum_{k=0}^{n-1} (P^k - P^*) \quad (n \geq 1) \quad (2.3)$$

としたとき, $H_n f$ が well defined で $\{H_n f\}$ が
ある関数 g に Cesaro 総和可能であるとき。これを
 $g = H f$ と表わし, g を charge f の P に関する potential
という。すなわち

$$g = H f = \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n H_k f \quad (2.4)$$

とくに S が finite ならば $I - P + P^*$ が正則
(Blackwell [2]) であるから

$$H = [I - P + P^*]^{-1} - P^* \quad (2.5)$$

となる。また P が single chain で transient ならば

$$H = [I - P]^{-1} \quad (2.6)$$

で, 従来の potential と一致している。この (2.3) で
定めた potential kernel H のもつ次の性質が, 後の議論
では有効な働きをする。

定理 1 関数 f, g において $P^* g < \infty$,

$$f = (I - P) g \quad (2.7)$$

とする。そのとき $H f < \infty$ で

$$g = H f + h \quad (2.8)$$

ここで h は regular で $h = P^* g$

(証明) (2.7) と H_n の定義 (2.3) から

$$H_n f = g - P^{n+1} g \quad (n \geq 1) \quad (2.9)$$

$P^*g < \infty$ であるから $\{P^n g\}$ は cesaro 総和可能。

したがって (2.9) より $\{H_n f\}$ もそうである $Hf < \infty$

$$Hf = g - P^*g \quad (2.10)$$

を得る。ここで $h = P^*g$ とおくと (2.10) は (2.8) と同じ式である。 ▮

この定理から Riesz 分解—super regular function—is non negative charge と regular function に一意に分解される—も成り立つことがわかる。

定理 2 (1) 関数 f が "non negative charge" かつ $P^*f = 0$ ならば "potential"

$$Hf \geq 0 \quad (2.11)$$

(2) f が "regular" ならば

$$Hf = 0 \quad (2.12)$$

(証明) (1) $H_n f = \sum_{k=0}^{n-1} P^k f$, $f \geq 0$ より $H_n f \geq 0$ である。 n について単調増加である $Hf < \infty$ から (2.11) を得る。

(2) すべての n に対して $f = P^*f = P^n f$ であるから $H_n f = 0$ (2.12) が成り立つことは明らか。 ▮

とくに P が transient であるならば $P^* = 0$ である。 (1) は potential H が pure であることを示している。

3. Theorems この § では policy improvement に用いる一般的な定理を述べる。

定義 2 P, \tilde{P} を 2 つの Markov 行列とし, $\{r_n(P)\}_1^\infty$ を P によって定まるあらかじめ与えられた S 上の有界関数列とする。

(1) $\{w_n = w_n(P)\}, \{u_n = u_n(\tilde{P})\}$ は次を満たすとする。

$$(I - P) w_{n+1} + P w_n = r_{n+1}(P) \quad n \geq 0 \quad (3.1)$$

$$(I - \tilde{P}) u_{n+1} + \tilde{P} u_n = r_{n+1}(\tilde{P}) \quad n \geq 0 \quad (3.2)$$

ただし $w_0 = u_0 = 0$ とする。

(2) 上の (1) で定めた $\{u_n\}, \{w_n\}$ に対して ($n \geq 1$)

$$f_n = r_n(P) - (I - P) u_n - P u_{n-1} \quad (3.3)$$

$$g_n = w_n - u_n \quad (3.4)$$

次の § で, これらの具体的な意味ははっきりするから, いまの段階では簡単にふれておく。この $\{w_n\}, \{u_n\}$ は § 1 での (1) ~ (4) などの case に対する MDP における reward で, それぞれ stationary policy に対応しているものである。しかし列の中で, 1 つだけが目的の reward であって, それ以外は補助的な reward の役割をしている。このパラメータ n は時間ではなく 1 つの MDP における reward の order のようなものを表わしている。(3.4) はしたがって reward の increment となる。また (3.3) は

reward $\{u_n\}$ に対する policy improvement の increment である。この $\{f_n\}$ の形は policy を perturb すれば得られるが、通常の Dynamic Programming や Linear programming にも関連している。MDP の policy improvement すべてにわたる原則は g_n すなわち reward の increment を正にすることであるが、それは f_n の符号に定まるということが本質である。§4 で述べることにしておき、そのための準備としていくつかの定理を用意する。

補題1 (1) $\{w_n\}$ は (3.1) の再帰式を満たすとする。

任意の関数列 $\{u_n\}$ ((3.2) を満たさなくてもよい) に対して

(3.3), (3.4) で $\{f_n\}, \{g_n\}$ を定めれば

$$f_n = (I - P)g_n + P g_{n-1} \quad (n \geq 1) \quad (3.5)$$

$$f_0 = 0 \quad g_0 = 0$$

(2) もし $P^*g_n < \infty$ ならば $H(f_n - P g_{n-1}) < \infty$ で

$$g_n = H(f_n - P g_{n-1}) + P^*g_n \quad (n \geq 1) \quad (3.6)$$

さらに $P^*g_{n+1} < \infty$ ならば $P^*f_{n+1} < \infty$ で

$$g_n = H(f_n - P g_{n-1}) + P^*f_{n+1} \quad (3.7)$$

$$P^*g_{n-1} = P^*f_n \quad (3.8)$$

(証明) (1) 定義 2 (2) において u_n を消去すればよい。

w_n が (3.1) をみたすことを用いれば (3.5) が成り立つ。

(2) $f = f_n - P g_{n-1}$, $g = g_n$ において 定理 1 を適用。■

この補題で P に関して g_{n-1} ($n \geq 1$) が "regular" であるとするれば, Pg_{n-1} も regular となるから, 定理 2(2) によって (3.7) は

$$g_n = Hf_n + P^*f_{n+1} \quad (3.9)$$

となる。したがって (3.9) を解釈すれば

「reward increment g_n は policy improvement の increment f_n を charge とする potential Hf_n と regular function P^*f_{n+1} との和に書き表わされる。」

もし P が transient ならば $P^*=0$ で (3.9) は

$$g_n = Hf_n \quad (3.10)$$

となり, 「policy improvement の increment を charge とする potential」(渡辺浩[4])で表わされる。 $f_n \geq 0$ ならば (2.11) より $g_n \geq 0$ で, これは Howard の policy improvement の単調性を示している。transient でないときは, $f_n \geq 0$ であっても $g_n \geq 0$ とはならない。しかし, 定理 3 で述べるように, $\{f_n\}$ のうちで f_{n-1}, f_n の 2 つの正負が g_{n-2}, g_{n-1} の正負を決定する。これは §1 の (2), (4) に対する原則で, さきの transient のときが (1) に対応している。(3) については特別で, g_n の正負も結論とするので, さらに仮定が必要である。

いま次の仮定を設ける。以下の定理はことわらない限

りこれを仮定するものとする。

仮定1 Markov 行列 P について state space S は

$$S = \bigcup_{k=1}^{\infty} R_k \cup T \quad (3.11)$$

で表わされ, 各 R_k は finite recurrent class, T は transient class とする。

S が有限ならばこれはつねに成り立っている。

与えられた P と $\{r_n\}$ によって $\{f_n\}$, $\{g_n\}$ を (3.3) (3.4) で定めたとする。

定理3 $\{w_n\}$ が (3.1) を満たし $n \geq 4$ で $P^* g_{\lambda} < \infty$ $\lambda = 1, \dots, n-1$ とする。

$$(f_1, \dots, f_{n-2}) = 0 \quad (3.12)$$

$$(f_{n-1}, f_n) \stackrel{\ell}{\geq} 0 \quad (3.13)$$

ならば次が成り立つ

$$(g_1, \dots, g_{n-3}) = 0 \quad (3.14)$$

$$(g_{n-2}, g_{n-1}) \stackrel{\ell}{\geq} 0 \quad (3.15)$$

(証明) (3.12) から (3.14) を帰納法で導く。 (3.9) より

$g_1 = H f_1 + P^* f_2$ だから $(f_1, f_2) = 0$ ならば $g_1 = 0$ なることは明らか。 (3.12) のとき $g_{n-3} = 0$ を示せばよい。

$$(3.7) \text{ より } g_{n-3} = H(f_{n-3} - P g_{n-4}) + P^* f_{n-2}$$

で, 帰納法の仮定によって $g_{n-4} = 0$ と $f_{n-3} = f_{n-2} = 0$ から $g_{n-3} = 0$ を得, (3.14) が示される。

次に (3.15) を証明する。 (3.12), (3.14) より (3.7) は

$$g_{n-2} = P^* f_{n-1} \quad (3.16)$$

の形。これは (2.2) より regular であるから (2.12) を用いると

$$g_{n-1} = H f_{n-1} + P^* f_n \quad (3.17)$$

を得る。 (3.13) より $f_{n-1} \geq 0$ だから (3.16) によって

$g_{n-2} \geq 0$ は明らか。したがって (3.15) を示すには

各 $\lambda \in S$ に対し $g_{n-2}(\lambda) = 0$ ならば $g_{n-1}(\lambda) \geq 0$ をいえばよい。ここで (3.11) の仮定を用いる。

1) $\lambda \in R_K$ のとき

P^* の $\lambda, j \in R_K$ 要素は positive である、いま

$$g_{n-2}(\lambda) = \sum_{j \in R_K} P_{\lambda j}^* f_{n-1}(j) = 0$$

だから $f_{n-1}(j) = 0 \quad j \in R_K$ によって (3.13) から

$f_n(j) \geq 0 \quad j \in R_K$ 。 (3.17) から

$$\begin{aligned} g_{n-1}(\lambda) &= \sum_{j \in R_K} H_{\lambda j} f_{n-1}(j) + \sum_{j \in R_K} P_{\lambda j}^* f_n(j) \\ &= \sum_{j \in R_K} P_{\lambda j}^* f_n(j) \geq 0 \end{aligned}$$

2) $\lambda \in T$ のとき

(3.16) によって $\sum_{j \in S} P_{\lambda j}^* f_{n-1}(j) = 0$ 。 (3.13) から

$f_{n-1} \geq 0$ だから 定理 2 (1) より

$$\sum_{j \in S} H_{\lambda j} f_{n-1}(j) \geq 0 \quad (3.18)$$

また $P_{\lambda j}^* > 0 \quad j \in \bigcup_k R_k$ より

$$f_{n-1}(j) = 0 \quad j \in \bigcup_k R_k$$

(3.13) より

$$f_n(j) \geq 0 \quad j \in \bigcup_k R_k$$

これから

$$\sum_{j \in \bigcup_k R_k} P_{\lambda j}^* f_n(j) \geq 0 \quad (3.19)$$

$\lambda \in T$ に対して (3.17) は

$$g_{n-1}(\lambda) = \sum_{j \in S} H_{\lambda j} f_{n-1}(j) + \sum_{j \in \bigcup_k R_k} P_{\lambda j}^* f_n(j)$$

であるから (3.18), (3.19) より $g_{n-1}(\lambda) \geq 0$ \square

系1 $(f_1, f_2) \leq 0$ ならば $g_1 \geq 0$

(証明) $P^*f_1 = 0, g_1 = Hf_1 + P^*f_2$ より同様に証明できる。 \square

系2 P について仮定1を以て $P^*g_1 < \infty, P^*f_2 < \infty$ とする。

$f_1 \geq 0$ かつ $f_2 \geq 0$ ならば $g_1 \geq 0$

(証明) 系1の証明中の関係式と定理2(1)より明らか。 \square

定理4 定理3につけ加え, $f_{n-1}(\lambda) = f_n(\lambda) = 0$ なる $\lambda \in S$ に対しては $P_{\lambda j} = \tilde{P}_{\lambda j} \quad \forall j \in S$ とする。もし

$g_{n-2} = 0$ ならば $(g_{n-1}, g_n) \leq 0$

(証明) $g_{n-2} = 0$ とすれば (3.7), (3.8) より

$$P^*f_{n-1} = 0 \quad (3.20)$$

$$g_{n-1} = Hf_{n-1} + P^*f_n \quad (3.21)$$

$$g_n = H[f_n - P g_{n-1}] + P^* g_n \quad (3.22)$$

を得る。(3.15)において $g_{n-1} \geq 0$ であるから
 $g_{n-1}(i) = 0$ のとき $g_n(i) \geq 0$ を示せば十分である。
 $f_{n-1} \geq 0$ ぞ (3.20) より

$$f_{n-1}(i) = 0 \quad i \in R = \bigcup_{R_1}^{\infty} R_R \quad (3.23)$$

も成りたっている。もし

1) $i \in R$ ならば

$$\begin{aligned} g_{n-1}(i) &= \sum_R H_{ij} f_{n-1}(j) + \sum_R P_{ij}^* f_n(j) \\ &= \sum_R P_{ij}^* f_n(j) \end{aligned}$$

$g_{n-1}(i) = 0$ としていえるから

$$f_n(j) = 0 \quad j \in R \quad (3.24)$$

(3.22)において (3.24) を用いれば

$$\begin{aligned} g_n(i) &= \sum_R H_{ij} \{f_n(j) - [P g_{n-1}]_j\} + \sum_R P_{ij}^* g_n(j) \\ &= \sum_R P_{ij}^* g_n(j) \end{aligned}$$

(3.23), (3.24) であるから 仮定(定理4)をみたす。

したがって この $i \in S$ に対しては $P_{ij} = \tilde{P}_{ij} \quad \forall j \in S$

このときには 明らかに

$$u_n(j) = w_n(j) \quad j \in R$$

したがって $g_n(j) = u_n(j) - w_n(j) = 0 \quad j \in R$

2) $i \in T$ ならば (3.21)において (3.23) を用いると

$$g_{n-1}(i) = \sum_T H_{ij} f_{n-1}(j) + \sum_R P_{ij}^* f_n(j)$$

ここで $H_{ij} > 0 \quad j \in T \cap T_0^c, H_{ij} = 0 \quad j \in T \cap T_0$

$P_{ij}^* > 0 \quad j \in R \cap R_0^c, P_{ij}^* = 0 \quad j \in R \cap R_0$ とすれば

$$f_{n-1}(j) = 0 \quad j \in T \cap T_0^c \quad (3.25)$$

$$f_n(j) = 0 \quad j \in R \cap R_0^c \quad (3.26)$$

(3.23), (3.25) より $f_{n-1}(j) = 0 \quad j \in R \cup T_0^c$ として (3.25) (3.26)

を用い, R_0, T_0 の定義方から

$$g_{n-1} = H f_{n-1} + P^* f_n = 0$$

1) の結果を用いれば $g_{n-1}(j) = 0 \quad j \in R$ より

$$g_n(j) = 0 \quad j \in R \quad (3.27)$$

を得る。これから $\lambda \in T$ に対する (3.22) は

$$g_n(\lambda) = \sum_{R \cap R_0} H_{ij} f_n(j) + \sum_{T \cap T_0^c} H_{ij} f_n(j)$$

となる。

$\sum_{R \cap R_0} P_{ij}^* f_n(j) = 0$, $f_n(j) \geq 0 \quad j \in R_0 \cup T_0^c$ でありか
ら, 定理 2 (1) より $g_n(\lambda) \geq 0$ となり, これが求

めり結果である。■

P に制限を加えれば

定理 5 Markov 行列 P は transient とする。

$(f_1, \dots, f_{n-1}) = 0$, $f_n \geq 0 \quad (n \geq 2)$ ならば

$$g_{n-1} = 0, \quad g_n \geq 0$$

(証明) (3.10) より明らか。■

4. Policy improvement of MDP 一般的な記号を定め, 各論に分けて §1 の (1) ~ (4) を論ずる。

S を state space として finite または denumerable。

Action space を表に出さず transition probability でおきかえる。つまり policy の定義を

$$\underline{P} = (P_1, P_2, \dots) \quad (4.1)$$

ただし $P_i, i=1, 2, \dots$ は $S \times S$ 上の Markov 行列とする。

\underline{P} の全体を Π と表わす。Markov 行列 P によって定まる S 上の有界関数を r_P とするとき

$$\underline{P} = (P_1, P_2, \dots) \in \Pi \text{ に対して}$$

$$\underline{P}^{(n)} r = P_1 \dots P_n r_{P_{n+1}} \quad n \geq 1 \quad (4.2)$$

と書き, これを n 期の expected reward という。ただし

$$\underline{P}^{(0)} r = r. \quad \text{この (4.2) の関数列をいかに total するかにより}$$

§1 の (1) ~ (4) の case が考えられている。

4.1. Discounted case $\underline{P}, \hat{\underline{P}} \in \Pi$ とする。

$$v_D(\underline{P}) = \lim_{n \rightarrow \infty} \sum_{k=0}^n \beta^k r_{P_k} \quad (4.3)$$

などとするとき, \underline{P} の方が $\hat{\underline{P}}$ より better とは

$$v_D(\underline{P}) \geq v_D(\hat{\underline{P}}) \quad (4.4)$$

(4.4) がすべての $\hat{\underline{P}} \in \Pi$ となりたつとき \underline{P} を discount optimal という。

定理 6 $v_D(\hat{\underline{P}}) < \infty$ とする。Markov 行列 P に対し

$$r_p + P v_D(\tilde{P}) - v_D(\hat{P}) \geq 0 \quad (4.5)$$

ならば stationary policy $\underline{P} = (P, P, \dots)$ の方が \hat{P} より better である。

(証明) 定義 2 で $r_n(P) = r_p$ $n=2$, $n \neq 2$ では $r_n(P) = 0$ とおく。 また $u_1 = w_1 = 0$, $u_2 = v_D(\tilde{P})$, $w_2 = v_D(\underline{P})$ とする。 こうすれば (3.3) より

$$f_1 = 0, \quad f_2 = r_p + P v_D(\tilde{P}) - v_D(\hat{P})$$

(4.5) から $f_2 \geq 0$ であるから 定理 5 によって

$$g_1 = 0, \quad g_2 = v_D(\underline{P}) - v_D(\hat{P}) \geq 0$$

よって (4.4) がなりたつ。 \square

4.2. Average case $\underline{P} \in \Pi$ に対し

$$v_A(\underline{P}) = \liminf_{n \rightarrow \infty} \frac{1}{n+1} \sum_{k=0}^n P^{(k)} r \quad (4.6)$$

$$v_A^{(2)}(\underline{P}) = \liminf_n \frac{1}{n+1} \sum_{m=0}^n \sum_{k=0}^m (P^{(k)} r - v_A(\underline{P})) \quad (4.7)$$

とする。 $\underline{P}, \tilde{P} \in \Pi$ に対し

$$v_A(\underline{P}) \geq v_A(\tilde{P}) \quad (4.8)$$

のとき \underline{P} の方が better といい, すべての \tilde{P} に対してなりたつとき \underline{P} を average optimal という。

定理 7 $v_A(\hat{P}), v_A^{(2)}(\tilde{P}) < \infty$ とする。

$$P v_A(\tilde{P}) - v_A(\hat{P}) \geq 0 \quad (4.9)$$

または $P v_A(\tilde{P}) - v_A(\hat{P}) = 0, r_p + P v_A^{(2)}(\tilde{P}) - v_A(\hat{P}) - v_A^{(2)}(\tilde{P}) \geq 0$ (4.10)

ならば $\underline{P} = (P, P, \dots)$ のほうが $\tilde{\underline{P}}$ より better つまり (4.8) が成り立つ。

(証明) $r_n(P) = r_P$, $n=2$, $r_n(P) = 0$, $n \neq 2$ とし
 $u_1 = v_A(\tilde{\underline{P}})$, $w_1 = v_A(\underline{P})$, $u_2 = v_A^{(2)}(\tilde{\underline{P}})$, $w_2 = v_A^{(2)}(\underline{P})$ とお
 けば (3.3) より

$$g_1 = v_A(\underline{P}) - v_A(\tilde{\underline{P}}) \quad (4.11)$$

$$f_1 = P v_A(\tilde{\underline{P}}) - v_A(\tilde{\underline{P}}) \quad (4.12)$$

$$f_2 = r_P + P v_A^{(2)}(\tilde{\underline{P}}) - v_A(\tilde{\underline{P}}) - v_A^{(2)}(\tilde{\underline{P}}) \quad (4.13)$$

となる。ここで定理3または系1によって $g_1 \geq 0$ となり (4.8) を得る。■

[注意] 定理7の(4.9), (4.10)の条件は $(f_1, f_2) \leq 0$ より強い。これは Derman [5] を denumerable にいいかえたからである。(4.9)と(4.10)の後半が同時になりたてば系2より $g_1 \geq 0$ となる (Aso and Kimura [1]) が $P^* g_1$, $P^* f_2 < \infty$ または $v_A(\tilde{\underline{P}})$, $v_A^{(2)}(\tilde{\underline{P}})$ の有界性が必要となる。

4.3. Nearly optimal case 以下では stationary policy $\underline{P} = (P, P, \dots)$ $\tilde{\underline{P}} = (\tilde{P}, \tilde{P}, \dots)$ を考える。 $0 < \beta < 1$ と

$$v_D(\beta, \underline{P}) = \lim_{n \rightarrow \infty} \sum_{k=0}^n \beta^k P^k r_P \quad (4.14)$$

とする。 \underline{P} のほうが $\tilde{\underline{P}}$ より better であるとは

$$\lim_{\beta \rightarrow 1} \{v_D(\beta, \underline{P}) - v_D(\beta, \tilde{\underline{P}})\} \geq 0 \quad (4.15)$$

\underline{P} が nearly optimal とは

$$\lim_{\beta \rightarrow 1} \{v_D(\beta, \underline{P}) - U(\beta)\} = 0 \quad (4.16)$$

ただし $U(\beta) = \sup \{v_D(\beta, \tilde{P}) : \tilde{P} \in \Pi\}$

Markov 行列 \tilde{P} に対して cesaro 和を \tilde{P}^* , potential kernel を \tilde{H} とするとき, 仮定 1 のもとでは $\tilde{P}^* r_{\tilde{P}}, \tilde{H} r_{\tilde{P}} < \infty$ と仮定から

$$u_1 = \tilde{P}^* r_{\tilde{P}}, u_2 = \tilde{H} r_{\tilde{P}} \quad (4.17)$$

とおく。この u_1, u_2 は (3.2) をみたす。いま

$$P u_1 \geq u_1 \quad (4.18)$$

$$r(P) + P u_2 \geq P u_1 + u_2 \quad (4.19)$$

を考える。

$G(\tilde{P}) = \{P : (4.18) \text{ が成立, } (4.18) \text{ が等号で } (4.19) \text{ が成立}$

または (4.18)(4.19) が等号で その $i \in S$ に対しては

$$P_{\lambda, j} = \tilde{P}_{\lambda, j} \quad \forall j \in S \}$$

とすれば

定理 8 $P \in G(\tilde{P})$ ならば $\underline{P} = (P, P, \dots)$ は $\tilde{P} = (\tilde{P}, \dots)$

よりも better。 $G(P) = \emptyset$ ならば \underline{P} は nearly optimal。

(証明) w_1, w_2 を P に対して (4.17) と同様に定め。

$$v_D(\beta, \underline{P}) = \frac{1}{1-\beta} w_1 + w_2 + o(1-\beta) \quad \text{と展開できるから,}$$

$$(f_1, f_2) \geq 0 \quad \text{と} \quad \lim \{v_D(\beta, \underline{P}) - v_D(\beta, \tilde{P})\} \geq 0$$

とは同値。したがって 定理 4 から得られる。 \square

この定理は S が finite のとき Veinott [12] が得たものを denumerable へと拡張している。ただし denumerable であるので、収束の条件が必要となって、仮定1が必要である。もちろん この仮定1は finite のとき自動的に満たされ、収束などの議論は気にしなくてよい。

4.4. Sensitive discounted case Discount case での Blackwell [2] による (1), (3) の結果を糸口として Miller, Veinott 等は (4.14) を $1-\beta$ で Laurant 展開すること考えた。これが sensitive discounted case であって 次のように定義される。

Stationary policy \underline{P} が $\tilde{\underline{P}}$ より n th order で better とは

$$\lim_{\beta \rightarrow 1} (1-\beta)^{-n} \{v_D(\beta, \underline{P}) - v_D(\beta, \tilde{\underline{P}})\} \geq 0 \quad (4.20)$$

また n th order で optimal とは (4.20) がすべての $\tilde{\underline{P}}$ に対して成り立つときをいう。

$$w_1 = P^* r_P \quad (4.21)$$

$$w_n = (-1)^n H^{n-1} P r_P \quad (n \geq 2) \quad (4.22)$$

と定義する。 u_n は P のかわりに \tilde{P} に対して定める。

$$g_n = w_n - u_n \quad (n \geq 1) \quad (4.23)$$

$$f_n = r_n(P) - (I-P)u_n - P u_{n-1} \quad (n \geq 1) \quad (4.24)$$

ただし $r_n(P) = r_p, n=2, r_n(P)=0, n \neq 2, u_0=0$

定理 9 $(f_1, \dots, f_{n-2})=0, (f_{n-1}, f_n) \leq 0$
 ならば stationary policy $\underline{P}=(P, P, \dots)$ は $\tilde{\underline{P}}=(\tilde{P}, \tilde{P}, \dots)$ より $(n-1)$ th order で better, つまり

$$\lim_{\beta \rightarrow 1} (1-\beta)^{-n+1} \{v_D(\beta, \underline{P}) - v_D(\beta, \tilde{\underline{P}})\} \geq 0 \quad (4.25)$$

が成立。

(証明) (4.21), (4.22) など で定めた $\{u_n\}, \{w_n\}$ は (3.1), (3.2) を満たすことがわかる。したがって (4.23) (4.24) で定めた $\{f_n\}, \{g_n\}$ は (3.3), (3.4) と同じものである。よって 定理 9 を示すには

$$(g_1, \dots, g_{n-1}) \leq 0 \quad (4.26)$$

と (4.25) が同値であれば 定理 3 より明らか。その同値であることは

$$v_D(\beta, \underline{P}) = (1-\beta)w_1 + w_2 + (1-\beta)w_3 + \dots \\
\quad \dots + (1-\beta)^N w_{n+2} + o(1-\beta)^{N+1} \quad (N \gg n) \quad (4.27)$$

となるから。(4.27) が成り立つことは (3.1) の再帰式を使って証明する。 ▮

この定理と同様な結果は "average overtaking optimal" な case についても 成り立つが, ここでは省略する。

参考文献

- [1] Aso, H. and M. Kimura ; An application of Markov potential theory to Markov decision processes, INT. J. System sci. (1973)
- [2] Blackwell, D. ; Discrete dynamic programming, AMS (1962)
- [3] ——— ; Discounted dynamic programming, AMS (1965)
- [4] Derman, C ; Markov sequential control processes — denumerable state space, J. Math. Anal. Appl. (1965)
- [5] ——— ; Finite state Markovian decision processes, Academic Press (1970)
- [6] Furukawa, N. ; Markov decision processes with compact action spaces, AMS (1972)
- [7] Howard, R.A. ; Dynamic programming and Markov processes, Wiley (1960)
- [8] Hordijk, A. ; Dynamic programming and Markov potential theory, Mathematical Center Tracts (1974)
- [9] Kemeny, J.G., J.L. Snell and A.W. Knapp ; Denumerable Markov chains, Van Nostrand (1966)
- [10] Miller, B.L. and A.F. Veinott, Jr. ; Discrete dynamic

programming with a small interest rate, AMS (1969)

- [11] Schaufele, R.A. ; A potential theoretic proof of a theorem of Derman and Veinott, AMS (1967)

- [12] Veinott, A.F. JR. ; On finding optimal policies in discounted dynamic programming with no discounting AMS (1966)

- [13] ——— ; Discrete dynamic programming with sensitive discount optimality criteria, AMS (1969)

- [14] 渡辺 浩 ; マルコフ計画法とポテンシャル,
日科技連計画シンポジウム (1967)